

**PIPELINED PARALLEL PROGRAMMING OPERATION IN A
NON-VOLATILE MEMORY SYSTEM**

5

ABSTRACT OF THE DISCLOSURE

10 The present invention allows for an increase in programming parallelism in a non-
volatile memory system without incurring additional data transfer latency. Data is
transferred from a controller to a first memory chip and a programming operation is
caused to begin. While that first memory chip is busy performing that program operation,
15 data is transferred from the controller to a second memory chip and a programming
operation is caused to begin in that chip. Data transfer can begin to the first memory chip
again once it has completed its programming operation even though the second chip is
still busy performing its program operation. In this manner high parallelism of
programming operation is achieved without incurring the latency cost of performing the
20 additional data transfers. Two sets of embodiments are presented, one that preserves the
host data in a buffer until successful programming of that data is confirmed and one that
does not require that success be achieved and that does not preserve the data thus
achieving a higher rate of data programming throughput.